Supporting Scholarly Research in Nautical Archaeology with Specialized Search Interfaces over Original Source Materials

Emma Carlson, Dr. Richard Furuta

***Abstract*—The project involves using mashups to ease the research process in nautical archaeology. The techniques used combine search results over the available resources and visualize the resulting collection. The implementation uses PHP to build a simple search interface over the New York Times, CNN, and JSTOR using their APIs. Visualization examples include text clouds and timelines.**

I. Background

THIS project explores using mashups to ease the research process for nautical archaeologists. Research is often time-consuming and frustrating for several reasons. First, typical search engines and online databases return many irrelevant results. Researchers are forced to waste time wading through articles which may have nothing to do with their particular topic. When the resource being investigated is something like a newspaper, with many different, unrelated articles per page, the relevance of a particular result may not be immediately apparent, and it may take quite a while to determine whether or not the result is useful. Second, it is impossible to search combined results from different databases. If researchers want to search different sources, they must perform their search through each individual source separately, then weed out irrelevant results, and somehow keep track of the relevant ones. Again, this is time consuming and can be complicated. Finally, from traditional search engines or databases, it can be difficult to gain a high-level understanding of the results which have been returned. With no way to visualize the body of results, important patterns might be missed.

Summer 2009 was spent developing a prototype intended to display how a few of these issues may be resolved, specifically how combining multiple resources and visualizing the results may be achieved.

II. Methods

Several questions were explored. To begin, what resources are available, and are they useful to nautical archaeologists? Many online databases were investigated, and a nautical archaeologist, Dr. Filipe Castro of the Department of Nautical Archaeology at Texas A&M University, was consulted. Dr. Castro explained that nautical archaeologists attempt to use shipwrecks to describe people's way of life. Because this is their main goal, nautical archaeologists are most interested in shipwrecks which occurred before people began taking records. Once people started keeping records, their way of life was well-documented, and so there would be very little left to learn about them by studying their shipwrecks. Information that is especially helpful to nautical archaeologists includes the cargo and nationality of ships, as well as when they were built or when they sank, as shown in Figure 1. This information is very helpful when trying to understand what everyday life was like around the time the ship sank, or when trying to understand, for example, how certain technologies spread across the world.

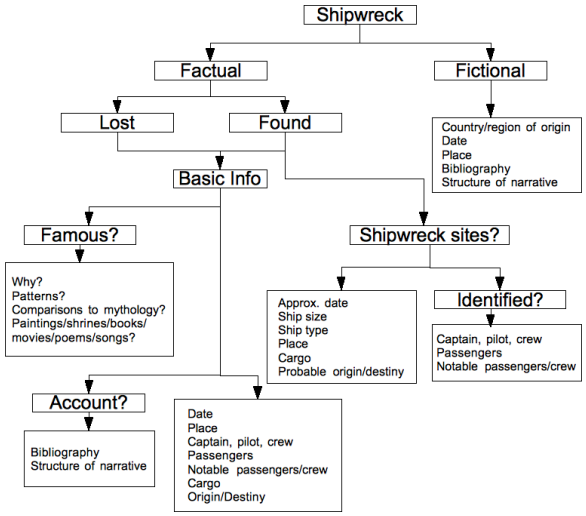

Fig. 1. Information nautical archaeologists would like to know about a given shipwreck.

With the interests of nautical archaeologists in mind, many online databases were explored. Several were chosen for the project based on the number of results they returned when queries like "shipwreck" were executed as well as what type of results they returned. Databases that returned fictional results (such as novels where characters are shipwrecked) were not considered to be particularly useful. The best databases discovered were JSTOR[5], Chronicling America: Library of Congress National Digital Newspaper Program[1],

Dr. Richard Furuta is with the Center for the Study of Digital Libraries and the Department of Computer Science and Engineering, Texas A&M University
Emma Carlson is with Harvey Mudd College.

This research was conducted as part of the CSE DREU 2009 program, which is supported by the NSF.

and the New York Times[6]. JSTOR has just under five million articles. A search for "shipwreck" yields a few thousand results, many of which are from archaeological journals and detail exactly the types of wrecks nautical archaeologists would be interested in. Chronicling America has scans and text versions of nearly a million pages of newspapers published between 1880 and 1910. The New York Times is recent news, but seemed like it might provide some useful information.

Next, what interfaces allow access to these resources' material? JSTOR's API uses a RESTful protocol called *Search and Retrieval by URL* (SRU)[5]. Requests are sent as URLs and search results are returned in XML. Ten results are returned at a time, with metadata including the author of the article, its title, the journal it appeared in, an abstract, subject words, and more. The New York Times Article Search API[6] provides results from 1981 to the present. Requests to the New York Times are also sent as URLs, but responses are returned in JavaScript Object Notation (JSON)[4]. The New York Times also returns ten responses at a time, with additional responses available if an "offset" parameter is added to the request URL. Chronicling America had planned to make an API available by Spring 2009, but it is not yet available. To provide a better model of working with multiple databases, CNN's articles via the Google Web Search API[3] were included instead of Chronicling America's.

Finally, what visualizations are useful over the resulting data? A text cloud plots word frequency by displaying more prominent words in larger text. This could help find patterns that otherwise might only be seen after reading a large number of articles. When studying a particular wreck, a timeline plotting the publication dates of articles might also be useful for exposing interesting trends.

Wordics[10] is a text cloud application which provides an API that is a block of JavaScript code allowing the user to specify colors, number of most frequent words displayed, and a URL from which to take text. MIT's SIMILE (Semantic Interoperability of Metadata and Information in unLike Environments)[7] project has a Timeline widget[9] which takes either XML or JSON data and plots points along a timeline. The Timeline API is also JavaScript code. The XML or JSON data provided to the Timeline can specify a title, date, URL, abstract, and other information.

The implementation uses PHP to build a simple search interface, shown in Figure 2. It is capable of searching the New York Times, CNN, and JSTOR through their respective APIs. As shown in Figure 3, responses in the form of URLs are sent to each respective database, and the responses are then converted. Results from the Article Search API and Google Web Search API, which both return responses in JSON, are converted into arrays using PHP's json_decode function. Results from JSTOR, which are in XML, are accessed using PHP's SimpleXML[8], which can convert XML to an object. A block of text comprised of the titles

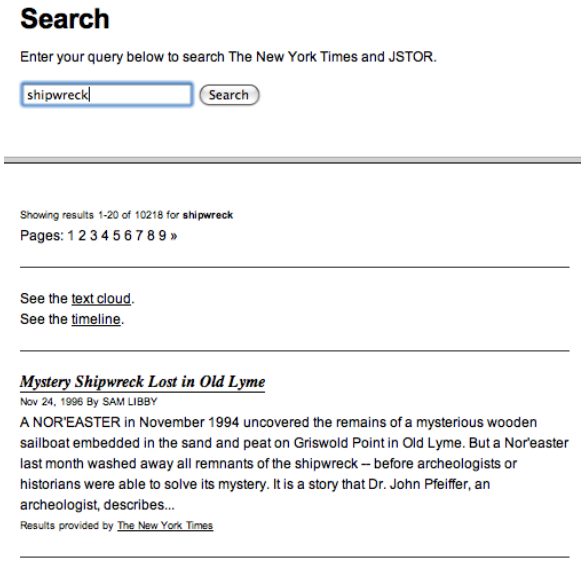

Fig. 2. Screenshot of the search interface with search query "shipwreck".

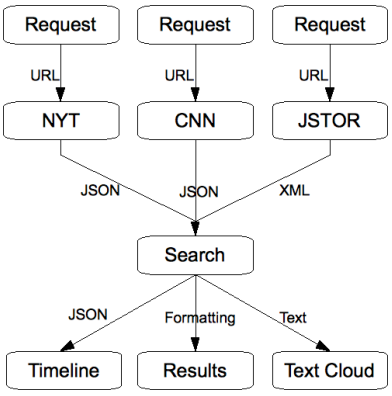

Fig. 3. The architecture of the interface.

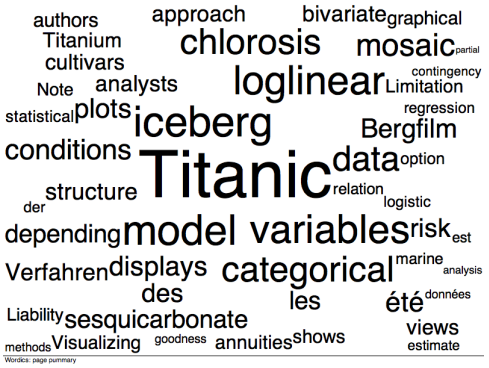

Fig. 4. Wordics text cloud showing the results for a search for "titanic".

and abstracts of the first 250 or so articles is sent to the text cloud, which is shown in Figure 4. Information such as the titles of articles, the dates they were published, their abstracts, and their URLs is encoded into JSON data which is then sent to the Timeline widget to be plotted, as shown in Figure 5. The text from the results is then formatted so it is more readable.

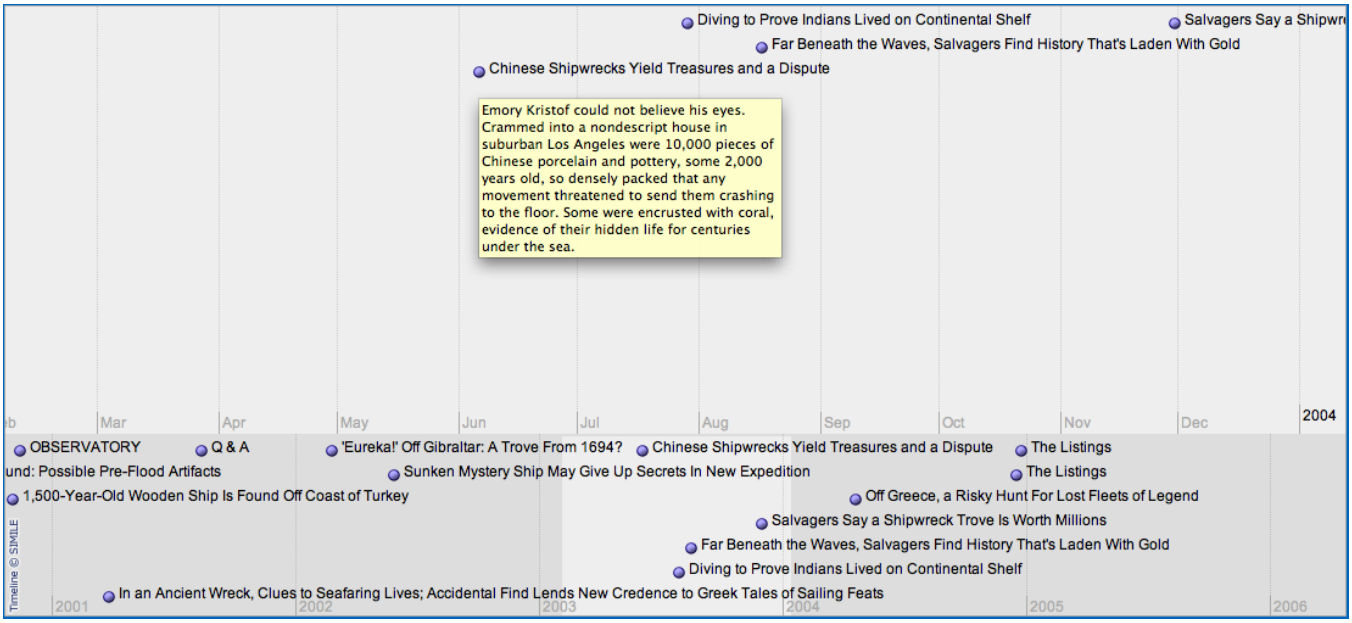

Fig. 5. Timeline showing the results of a search for “nautical archaeology”.

III. Results

The prototype currently works mostly as described in the Methods section. However, there are still a few issues. Wordics is not consistent and has several bugs. It is slow and does not always refresh the search term - that is, if you create two text clouds in a short amount of time over two different queries, sometimes Wordics will continue to return a text cloud of only the first query. The Timeline also has a few issues that are still being investigated. Not every search term is displayed. For example, if “titanic” is the keyword, results are yielded, but none of them are plotted on the timeline. By contrast, the results for “nautical archaeology” as the search query show up perfectly in the timeline.

Also, not all of the accessed databases returned the same information in their response. For example, JSTOR responses do not include a URL to access the particular article. This means that currently, the results returned by JSTOR in the search only provide a title, author, date, and abstract, with no link for the user to access the article. Hopefully in the future there will be a way to provide a link to the article. Also, the New York Times articles always provide a day, month, and year for the publication dates of their articles, while JSTOR is not as consistent. Some articles have no dates, or only years, et cetera. This means the timeline sometimes fails to plot JSTOR articles correctly, so currently the prototype has the timeline only using the New York Times articles.

IV. Discussion and Conclusion

JSTOR contains articles from archaeological journals, which would definitely be useful to nautical archaeologists. However, the New York Times and CNN give access to fairly recent (since the 1980’s) news. As discussed above, news this recent is not particularly useful to nautical archaeologists, so in the future, hopefully the search interface will be able to incorporate other resources which may be more relevant to nautical archaeology.

Also, the search currently has no specialization toward nautical topics. In the future, it would be useful to have the search recognize certain nautical terms - for example, a search for “ship” could also look at documents containing the words “boat”, “steamer”, or “schooner”. Along this same line, ideally the search would be able to distinguish between relevant and irrelevant articles. Many ships share their names with cities or even states, and conventional searches have no way of distinguishing a ship from a state. For example, a search for “valencia” would ideally return only articles concerned with the wreck of the steamship Valencia off the coast of Vancouver Island, and not articles mentioning Valencia St. or persons named Valencia.

Finally, it might be interesting to explore still more methods of visualization, such as the data visualizations available from Flare [2]. Flare offers complex visuals which can be animated and interactive. Several of their visualizations might be very useful when trying to better understand a collection of search results.

References

[1] Chronicling America: LIbrary of Congress National Digital Library Program. [Online] June 24, 2009. Available: http://chroniclingamerica.loc.gov/
[2] Flare: Data Visualization for the Web. [Online] August 6, 2009. Available: http://flare.prefuse.org/
[3] Google Web Search API. [Online] July 17, 2009. Available: http://code.google.com/apis/ajaxsearch
[4] JSON. [Online] August 6, 2009. Available: http://www.json.org/
[5] JSTOR. [Online] August 3, 2009. Available: http://dfr.jstor.org/

[6] The New York Times. [Online] July 10, 2009. Available: http://developer.nytimes.com
[7] SIMILE - Semantic Interoperability of Metadata and Information in unLike Environments. [Online] July 28, 2009. Available: http://simile.mit.edu/
[8] SimpleXML. [Online] August 3, 2009. Available: http://us2.php.net/simplexml
[9] Timeline Web Widget for Visualizing Temporal Data. [Online] July 28, 2009. Available: http://www.simile-widgets.org/timeline/
[10] Wordics. [Online] July 24, 2009. Available: http://wordics.lemmatica.com